IDPS-SDN-ML: An Intrusion Detection and Prevention System Using Software-Defined Networks and Machine Learning

Tamara AlMasri
Department of Computer Science/Cybersecurity
Princess Sumaya University for Technology (PSUT)
Jordan, Amman
tamaranidal.almasri@gmail.com

Mohammad Abu Snober
Department of Computer Science/Cybersecurity
Princess Sumaya University for Technology (PSUT)
Jordan, Amman
m.abusnober@psut.edu.jo

Qasem Abu Al-Haija
Department of Computer Science/Cybersecurity,
Princess Sumaya University for Technology (PSUT)
Jordan, Amman
q.abualahaija@psut.edu.jo

2022 International Conference on Smart Technology, Applied Informatics, and Engineering (APICS) | 978-1-6654-8985-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/APICS56469.2022.9918804

Abstract—The recent increase in the amount and precision of cyberattacks necessitates the development of detection and prevention systems that mitigate the risks of these threats. Several studies discussed using Software-Defined Networks (SDNs), Challenge-based collaborative intrusion detection systems, or pattern recognition using machine learning or machine learning. The Intrusion Detection and Prevention Systems (IDPS) monitor traffic to detect unusual traffic and compare the network traffic with known attacks to identify anomalies. As a response, the system alerts the network administrator or controller of the possible attack and blocks (prevents) the attack. It is crucial to detect any attack early to avoid huge damage to the network or the system. This study aims to suggest a new method that combines the pattern recognition of machine learning and the network programmability feature and architecture to better protect and defend the network against Denial of Service (DoS) and Port Scanning attacks. A machine learning algorithm was built using Anova for feature selection and applying the chosen features to multiple machine learning models. The naive Bayes machine learning model achieved the highest accuracy with 86.9% for DoS attacks and 93.5% for Probe attacks.

Keywords—Intrusion Detection and Prevention Systems (IDPS), Software-Defined Networks (SDN), Machine Learning (ML), Denial of Service (DoS) Attacks, Port Scanning Attacks, NSL- KDD, Anova, Naive Bayes.

I. INTRODUCTION

In the last decade, we have witnessed a huge augmentation in the numbers and sophistication of cyber-attacks against critical information, the infrastructure of individuals, and organizations. It is estimated that 4% of network attacks are port scanning and 10% are Denial of service (DoS) of the total number of attacks [1]. In addition, to threatenT) and cyber-physical systems (CPS) [2] to harm the reing the information and assets, several other attacks are targeting the connectivity of the Internet of things (Ioal-time communication services of IoT and CPS.

With this increment in cyber-attacks and their complexity, there is a major development and research of the countermeasures to mitigate these risks; security experts are exploring using Software- Defined Networking (SDN) technology for real-time and efficient defense against cyber threats. According to a study by Verizon [3], 15% of organizations have SDN implemented, while 57% are expected to implement SDN within the next two years. This requires early detection of attacks, typically using anomaly-based detection approaches. Earlier researchers addressed the issue of the rising threats and their complexity by analyzing network anomalies and creating the concept of Software-Defined networks, as in "SDSecurity: A Software-Defined Security experimental framework [4], which focuses on separating the data plane from the control plane by providing a flexible and centralized security solution. This is achieved by abstracting the security mechanisms from the hardware layer to a software layer. However, the paper only focused on one aspect of handling network anomalies & cyber threats.

Also, machine learning [5] and the possibility of implementing pattern recognition on network traffic and attack patterns have been considered for intrusion detection. For example, Atik Abubakar; Bernardi Pranggono [6] suggested an approach to detecting network intrusion to overcome the limitation of signature-based Intrusion Detection Systems (IDS) and provided intrusion detection for attacks using pattern recognition while using an existing intrusion detection software, such as Snort. Celyn Birkinshaw, Elpida Rouka, Vassilios G. Vassilakis [7]. Thought of a design proven to be extremely beneficial: the use of SDN together with the Credit Based Threshold Random Walk (CB-TRW) and PortBingo (PB) algorithms and the OpenFlow Protocol as an intrusion detection system to defend against PortScanning and DoS attacks. One of the drawbacks was the demonstrated huge load on the CPU usage and fully relying on the information provided by the controller.

This paper intends to suggest an intrusion detection and prevention framework that relies mainly on SDN and compatible machine learning algorithms to detect various types of port scans and DoS attacks which should provide sufficient protection and efficiency.

The rest of this paper is divided into sections: Section II introduces the background and related work. Section III presents IDS using SDN & Machine Learning methodology, Section IV the machine learning results, and the hypothesis expected results. Finally, Section V shows conclusions and highlights future works.

II. METHODOLOGY

The proposed system methodology can be decomposed into two main subsystems: the machine learning (ML)

978-1-6654-8985-0/22/$31.00 ©2022 IEEE

subsystem and the software-defined network (SDN) subsystem.

A. Machine Learning

1) Dataset: In this paper NSL-KDD [8] dataset was used. NSL-KDD is an enhanced version of the KDD99 dataset. NSL-KDD is meant to solve multiple issues in the KDD99 dataset, one of which is the large number of duplicates found in the KDD99 dataset. The NSL-KDD training dataset consists of 125,972 rows and 43 columns, while the test dataset contains 22,543 and 43 columns. The dataset contains four attack categories: DoS, probe, U2R [9], and R2L [10]; table 1 defines the mentioned attacks. In addition to subcategories of attacks under each corresponding category. It also explains the types of attacks defended in this paper. For this paper, we will be focusing on DoS and Probe (Port Scanning) attacks.

2) Data Preprocessing: The following was achieved in the data preprocessing step: Firstly, both training and test datasets did not contain column names [14] . Each column was given its correct name. Secondly, One-hot encoding-defined- [15] was used to transform any categorical-defined- [16] data into non-categorical - defined- [17] data as the dataset included both types. Thirdly, changing attack types from a string to a numerical representation: 0= normal traffic flow, 1= Dos attack, and 2= Probe attack. Lastly, feature scaling was performed, a technique used in machine learning to standardize the dataset such that all the data values are in a fixed range to avoid having features with large values that might impact the final results.

3) Feature Selection: After performing One-hot encoding, the output of the dataset resulted in a total of 123 columns instead of the original 42 columns. In this paper, three methodologies were tested:

- Building the model using all the features
- Building the model using features extracted by Anova, which determines whether the means from the 123 features come from the same distribution or not. Anova feature selection resulted in 13 features for each attack type [18]. Table 2 displays the features selected when Anova was used.
- Recursive feature elimination (RFE) [19] was also applied for feature selection; however, as RFE depends on the machine learning model as one of the input variables, after testing with the four chosen machine learning models, a dead-end was reached as not all machine learning models can be used as an input to feature selection. Hence for a fair comparison, RFE was disregarded.

4) Building the Model: This paper's dataset was trained and tested on four different models. Table 3. Each model was tested twice, first using all the features and second using ANOVA extracted features. The output from each model and classifier was compared to decide on the model that achieved the highest accuracy and the most suitable classifier.

TABLE I. ATTACK DEFINITIONS AND TYPES [13]

Attack	Definition	Attack Types
DoS	It is an attack meant to compromise a machine's or network's availability by making it inaccessible to users.[11]	"Back," land," "Neptune," "Pod," "Smurf," "Teardrop," "mailbomb," "apache2", "process blue," "udp-storm," "worm."
Probe	It is an attack that exploits the computer's vulnerabilities to obtain access to the system and its files.[12] .	"Ipsweep," "nmap," "portsweep," "satan," "mscan," "saint"

TABLE II. THE OUTPUT OF SELECTED FEATURES (ANOVA)

Dataset: Attack	Features
Training Dataset: DoS - ANOVA	"Logged_in","count","serror_rate", "flag_S0","Srv_serror_rate", "same_srv_rate","flag_SF" "Srv_diff_host_rate", "dst_host_count", "Dst_host_srv_count", "service_http", "Dst_host_same_srv_rate", "service_private", "Dst_host_srv_diff_host_rate", "Dst_host_serror_rate", "service_smtp", "dst_host_srv_serror_rate", "Protocol_type_udp", "service_domain_u"
Test Dataset: DoS - ANOVA	"logged_in","count","rerror_rate", "srv_rerror_rate","same_srv_rate", "dst_host_count","dst_host_same_srv_rate", "dst_host_rerror_rate", "dst_host_srv_rerror_rate","Service_http", "flag_REJ", "flag_SF,""dst_host_srv_count"
Training Dataset: Probe - ANOVA	"Logged_in" ,"rerror_rate", "flag_SF" "Srv_rerror_rate","dst_host_srv_count", "Dst_host_diff_srv_rate","dst_host_rerror_rat e" "Dst_host_srv_diff_host_rate", "dst_host_srv_rerror_rate "service_eco_i"","service_private", "Dst_host_same_src_port_rate", "Protocol_type_icmp"
Test Dataset: Probe - ANOVA	"Logged_in", "rerror_rate", "Srv_rerror_rate", "same_srv_rate", "Diff_srv_rate", "dst_host_srv_count", "Dst_host_same_srv_rate", "flag_SF" "Dst_host_diff_srv_rate", "flag_REJ", "Dst_host_rerror_rate", "service_http", "dst_host_srv_rerror_rate"

TABLE III. CLASSIFIERS DEFINITIONS [20]

Classifier	Definition
Decision Tree	A classifications algorithm builds a tree-structured model that branched into smaller subsets at each level.
Random Forest	A classification algorithm creates several decision trees at the training time and generates the class classification or regression for each tree separately.
Naive Bayes	A classification algorithm that assumes the independence of predictors as in Bayes Theorem.

B. Software Defined Networks

1) SDN Definition: A network architecture around the network programmability feature allows the network to act intelligently or be controlled centrally by embedded software applications. Which provides effective detection and monitoring for network security issues & attacks. SDN

focuses on separating the control plane from the data plane, allowing the network to block traffic when dangerous activity is detected instantly. Machine learning can help SDN-based Network Intrusion Detection Systems (NIDS) overcome network security vulnerabilities.

2) SDN Architecture: SDN system structure is based on the OpenFlow protocol, considered the most common SDN protocol as it is used to connect the controller and the network switches. The SDN-based IDPS is divided into two parts: The intrusion detection part, which includes firing an alert to the network controller to notify a user of the occurrence of a potential threat, and the intrusion prevention part, which automatically creates and sends a command to all data forwarding devices to drop paEach plane has its specification and role in the programmable network to improve its response and flexibility: SDN-Data plane (Infrastructure layer) consists of all physical and virtual network data forwarding devices. Fig. 2 shows SDN Data Plane. ckets from the attacker. SDN architecture Fig. 1 consists of the Data plane, Control plane, and Application plane.

SDN Controller is the most important component, and it is considered the system's brain because it controls both the data plane devices and the application layer functions. Also, it has its own Network Operating System and interfaces APIs. The SDN controller is the programmable part of the network in which most of the used network algorithms exist and all programmable APIs. The controller typically exists on a server to control the data flow and network policies. Fig. 3 SDN Controller.

SDN Application plane is the component that consists of the common network functions and applications, such as intrusion detection systems, firewalls, and many others. In this hypothesis, we are testing the possibility of employing a top-down approach [22] that makes use of multiple algorithms that aim to enhance and fix the regular SDN automity, such as Pre-tested intrusion detection algorithms (CB-Threshold and Port Bingo), Dynamic Traffic Scheduling algorithm [23][24] , Chaos-Genetic algorithm [25] and Admission Control algorithm [24]. These algorithms are structured to optimize network performance and mitigate DOS and Port scanning attacks. The SDN components communicate using application protocols: the infrastructure layer communicates with the control layer using the Southbound APIs - depending on the OpenFlow protocol- and the control layer communicates with the application layer using the Northbound APIs.

III. Results And Discussion

To calculate the accuracy achieved by each model, a confusion matrix was built, which displayed true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) [26] results.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

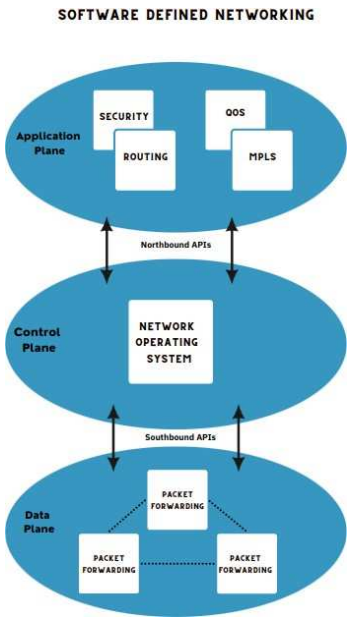

Fig. 1. SDN Architecture [21].

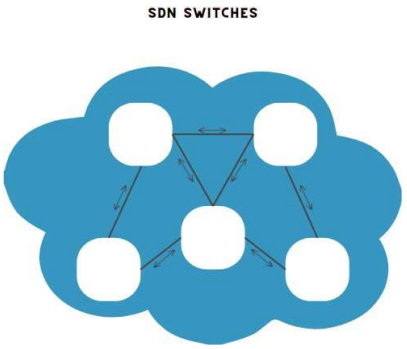

Fig. 2. SDN-Data plane [21].

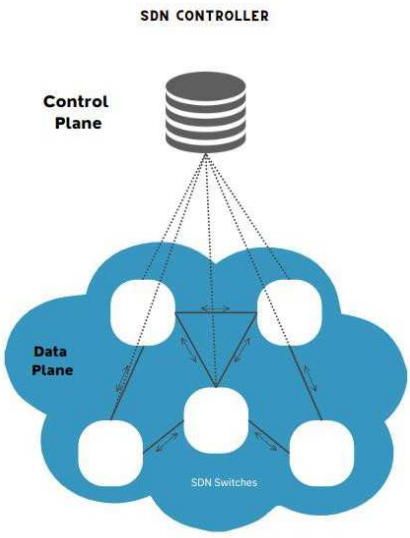

Fig. 3. SDN Controller [21].

After testing the dataset against four different models, it was observed that the Naive Bayes model had the highest accuracy when Anova features selection was applied. Below are the confusion matrices for DoS (Fig. 4) and Probe (Fig. 5) attacks using the Naive Bayes machine learning model, with 13 features extracted using Anova.

The accuracy of DoS using the Naive Bayes machine learning model and Anova features equals 86.9%. The accuracy achieved for Probe using the Naive Bayes machine learning model and Anova features equals 93.5%. Finally, with the provided results, an intrusion detection and prevention system will be built by combining the SDN architecture, the chosen machine learning model, feature selection model, and NSL-KDD dataset to reveal any abnormalities in the network.

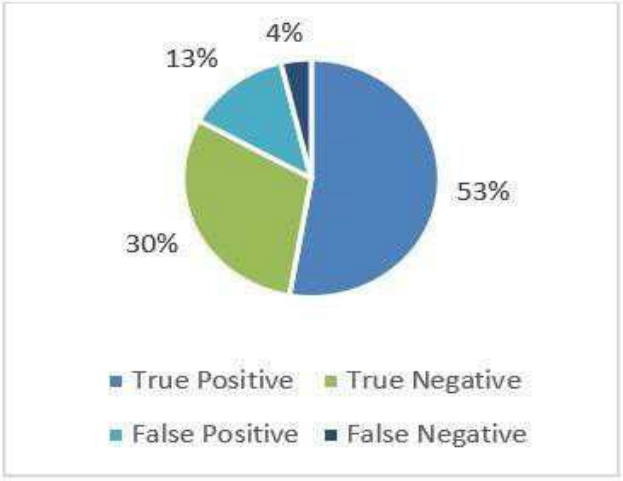

Fig. 4. Confusion matrix: DoS Attack

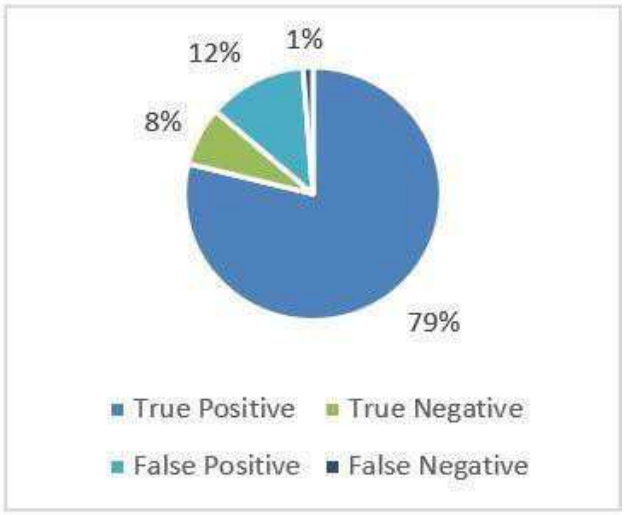

Fig. 5. Confusion matrix: Probe Attack

IV. Conclusion

This paper suggested a new approach to detect and prevent network intrusions by creating a machine-learning algorithm to detect DoS and Port Scanning attacks using the NSL-KDD dataset. Second, by combining the output of the machine learning algorithm with the demonstrated SDN architecture, its performance is expected to improve after utilizing multiple previously tested algorithms. One hot encoding was used to transform the categorical data to non-categorical data to perform the data preprocessing and the ANOVA test for feature selection. Multiple classifiers were tested on the original data set and the chosen features of the ANOVA test, and the best results were produced via the Naive Bayes model. The model resulted in an accuracy of 86.9% for DoS attacks and an accuracy of 93.5% for Port Scanning attacks. Previous works suffered performance issues in high-speed and high-load networks. The suggested approach is expected to overcome these issues with the limitations of signature-based IDS. In the future, the suggested approach claims should be tested using VMs multiple times to prove their efficiency and accuracy. The core structure of the SDN will be built upon Coursera's SDN by embedding the suggested algorithms.

References

[1] Q. A. Al-Haija, "On the Security of Cyber-Physical Systems Against Stochastic Cyber-Attacks Models," 2021 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-6, 2021.

[2] Q. A. Al-Haija, C. D. McCurry and S. Zein-Sabatto, "A Real Time Node Connectivity Algorithm for Synchronous Cyber Physical and IoT Network Systems," 2020 SoutheastCon, pp. 1-8, 2020.

[3] V. Lonker "Thinking beyond the box – how Software Defined Networks are changing the future of connectivity," Verizon News Archives, 2018.

[4] A. Darabseh, M. Al-Ayyoub, Y. Jararweh, E. Benkhelifa, M. Vouk, and A. Rindos, "SDSecurity: A Software-Defined Security experimental framework," 2015 IEEE International Conference on Communication Workshop (ICCW), London, 2015, pp. 1871-1876, DOI: 10.1109/ICCW.2015.7247453.

[5] Q. Abu Al-Haija, Al-Saraireh, J. Asymmetric Identification Model for Human-Robot Contacts via Supervised Learning. Symmetry 2022, 14, 591. https://doi.org/10.3390/sym14030591

[6] A. Abubakar and B. Pranggono, "Machine learning-based intrusion detection system for software define software-defined", 17th International Conference on Emerging Security Technologies (EST), Canterbury, pp. 138-143, 2017.

[7] Celyn Birkinshaw, Elpida Rouka, Vassilios G. Vassilakis, " Implementing an intrusion detection and prevention system using software-defined networking: Defending against port-scanning and denial-of-service attacks," Journal of Network and Computer Applications, Volume 136, 2019, Pages 71-85, ISSN 1084-8045.

[8] Abu Al-Haija, Q.; Zein-Sabatto, S. An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks. Electronics 2020, 9, 2152. https://doi.org/10.3390/electronics9122152

[9] D. Hassan "Cost-Sensitive Access Control for Detecting Remote to Local (R2L) and User to Root (U2R) Attacks". International Journal of Computer Trends and Technology (IJCTT) V43(2):124-129, 2017.

[10] Abu Al-Haija, Q.; Al-Badawi, A. Attack-Aware IoT Network Traffic Routing Leveraging Ensemble Learning. Sensors 2022, 22, 241. https://doi.org/10.3390/s22010241

[11] Abu Al-Haija, Q.; Al-Dala'ien, M. ELBA-IoT: An Ensemble Learning Model for Botnet Attack Detection in IoT Networks. J. Sens. Actuator Netw. 2022, 11, 18. https://doi.org/10.3390/jsan11010018

[12] Tech Target Contributor, "Probe", SearchSecurity- TechTarget, Circuit-switched services equipment and providers, Section 3, 2005.

[13] I. Cosimo, A. Ahsan, G. Mandar, D. Kia et. al. Statistical Analysis Driven Optimized Deep Learning System for Intrusion Detection. Defcom17, "NSL_KDD," Github, NSL_KDD, Field Names.csv, 27bbbdf, 2015.

[14] One Hot Encoding: One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Rakshithvasudev, "What is One Hot Encoding? Why and When Do You Have to Use it? " Hackernoon, 2017.

[15] Categorical Data: a collection of information that is divided into groups and can take on numerical values. Formplus, "Categorical Data: Definition + [Examples, Variables & Analysis] " , Formplus, 2007.

[16] Non-Categorical Data: as range variables and are expressed as numbers within the set of the Reals, usually varying from negative infinity to positive infinity. Benjamin Dickman, Central Connecticut State University (2014), " What are non-categorical variables in statistics? "Quora, 2019.

[17] Sampath Kumar Gajawada, "ANOVA for Feature Selection in Machine Learning," Towards data science, Applications of ANOVA in feature selection, 2019.

[18] Jason Brownlee, "Recursive Feature Elimination (RFE) for Feature Selection in Python," Machine learning mastery, making developers awesome at machine learning, data preparation, 2020.

[19] Mandy Sidana, " Intro to types of classification algorithms in Machine Learning, "Medium, Types of classification algorithms in machine learning, 2017.

[20] IPCisco, " SDN architecture components, "IPCisco, 2020.

[21] Al-Haija QA. Top-Down Machine Learning-Based Architecture for Cyberattacks Identification and Classification in IoT Communication Networks. Frontiers in Big Data, vol.4, 2021.

[22] Q. A. Al-Haija, E. Saleh and M. Alnabhan, "Detecting Port Scan Attacks Using Logistic Regression," 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2021, pp. 1-5, doi: 10.1109/ISAECT53699.2021.9668562.

[23] H. Ren, X. Li, J. Geng, and J. Yan, "An SDN-Based Dynamic Traffic Scheduling Algorithm," in the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chengdu, 2016 pp. 514-518. DOI: 10.1109/CyberC.2016.103

[24] Gharooni-fard G., Moein-darbari F. A New Approach to Network Optimization Using Chaos-Genetic Algorithms. In: Yang XS., Koziel S. (eds) Computational Optimization and Applications in Engineering and Industry. Studies in Computational Intelligence, Springer vol 359, 2011 .

[25] J. Leguay, L. Maggi, M. Draief, S. Paris and S. Chouvardas, "Admission control with online algorithms in SDN," NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium, Istanbul, 2016, pp.718-721,

[26] Al-Haija, Q.A.; Alsulami, A.A. High-Performance Classification Model to Identify Ransomware Payments for Heterogeneous Bitcoin Networks. Electronics 2021, 10, 2113.